Video Perception and Forecasting Models for Autonomous Systems

Sudhir Yarram

University at Buffalo

Modern AI systems have achieved remarkable success in processing and generating videos, yet they falter in their abilities to *perceive* and *forecast* the future. In contrast, an integral part of how humans and other biological entities interact with the world is our innate ability to comprehend and forecast, raising the critical question: *how do we bridge this gap?*

The success of today's machine learning approaches hinges on availability of large volumes of high-quality data. However, the challenge extends beyond merely accessing vast datasets; it requires efficient training methodologies and the development of models domain adaptable enough to transfer knowledge across different domains seamlessly. Additionally, most current methods for future forecasting fail to capitalize on the rich semantic cues derived from perception, leading to frameworks that struggle with the intrinsic complexity of predicting the future. Thus, building adaptable perception models that simplify forecasting tasks becomes vital for developing general-purpose autonomous systems. Achieving seamless integration and coexistence of autonomous systems within human environments—especially in complex applications like autonomous driving and augmented reality—calls for advancements across multiple fronts: novel algorithms for video perception, perception-informed future forecasting, and integration of agent intentions into the forecasting models.

My research focuses on designing video perception and forecasting models for autonomous systems that will enable autonomous systems to *perceive* and *forecast* in the real world just as humans are able, and ultimately surpass humans ability to forecast.

Building video perception models will enable more domain adaptable and efficient operation with significantly lesser computation resources and training data, by building efficient computational mechanisms and domain adaptation techniques. With these innovations in perception, we identify new, forecasting conducive scene representations that can model semantic-aware agent dynamics for better forecasting. Moreover, we aim to boost the forecasting controllability by integrating agent action intentions as language commands. These innovations are crucial for autonomous systems to comprehend their surroundings and accomplish tasks effectively.

This research primarily focuses on

1. **Video Perception** Developing efficient and domain adaptable models that enable autonomous agents to understand scene semantics from images and videos, which is essential for tackling diverse environments.

2. **Future Forecasting** Leveraging video perception to design models with enhanced abilities to predict future scenes, making use of depth information to encode 3D knowledge of the world.
3. **Multimodal Future Forecasting** Integrate vision, agent intentions as language, to create models that achieve far greater forecasting controllability than visual input alone.

0.1 Video Perception

We develop video perception models enabling autonomous systems to grasp the semantics of objects and scenes from images and videos. Of particular interest are domain adaptable and efficient models that are adaptable across domains and require less compute and training time to build. This effort addresses a significant challenge in AI: making autonomous systems capable of understanding complex environments in a manner that is both computationally efficient and broadly applicable.

Contributions.

Efficient Attention for video instance segmentation: In [1, 2], I have introduced efficient attention mechanism titled "deformable spatio-temporal attention mechanism" that can reduce the high computational cost of attention in transformer for video processing by 10 times while maintaining high performance. This mechanism computes attention for reference points from a small, fixed set of keys predicted by a learned model, allowing for linear rather than quadratic computation relative to the size of spatio-temporal feature maps. This innovation is the cornerstone of the first efficient Video Instance Segmentation system that addresses training inefficiencies in video-based transformer models, enabling faster iterations and the deployment of VIS models crucial for realizing easy-to-build autonomous systems.

Domain adaptation models for semantic segmentation: In a series of works [3, 4] I have developed domain adaptive models that exhibit robustness in target domains without annotations adapting trained models on source domain data with annotations. We have explored various mechanisms such as adversarial structured prediction, local-global alignment, to guide domain adaptation. While cross-domain perception is innate to humans, such behaviors are hard to replicate in computer vision systems, owing to vast variations in terms of lighting, texture, object appearances and backgrounds. In [4], we presented, for the first time, a adaptation mechanism that could leverage the structural similarities in the semantic segmentation outputs as supervision signal. Key to enabling this is building a regularizer and adversarial setup through which we make structured predictions across source and target domains follow the spatial layout learned from the source ground truth. Since the spatial layouts across domains share similarities, we can reliably align source and target features even without explicit

alignment. In [3], we propose another adaptation mechanism that explicitly focus on both global(image) level alignment and local level alignment of features across domains. This results in more fine grained alignment of features across domains, resulting in enhanced semantic segmentation performance in the target domain.

Future Directions.

Active perception A pivotal future direction is integrating active perception into our video perception models. Moving beyond the limitations of episodic mapping, this approach aims to equip autonomous systems with continuous environmental understanding. By actively influencing their perception through decisions, systems will achieve a dynamic understanding akin to biological entities, greatly enhancing their adaptability in complex settings.

0.2 Future Forecasting

Utilizing the insights gained from video perception, we aim to develop future forecasting models that offer far greater ability to forecast future scenes compared to future forecasting without leveraging such privileged information. Successful future forecasting requires solving intricate challenges such as estimating scene geometry, forecasting future motion, and synthesizing content from novel viewpoints. By leveraging the semantics captured in video perception beyond what RGB images alone can offer, we aim to disentangle the intertwined aspects of scene geometry and motion, and thereby increase efficacy of forecasting future scenes by enabling image rendering from novel perspectives.

Contributions.

Disentangled 3D scene representation The integration of scene geometry modeling, motion forecasting, and novel view synthesis is essential yet challenging for accurate future forecasting. Traditional methods often adopt an entangled approach, facing two primary limitations: the oversimplified assumption that scene geometry can be modeled using a layered approach suitable for homography-based warping and that future motions can be approximated with affine transformations. Additionally, this layered approach struggles to model future 3D motion accurately due to the complexity of simultaneously predicting ego-motion and the motion of dynamic objects. In my research, I have introduced the first disentangled 3D scene representation model. This model, enhanced by estimated depth, successfully separates scene geometry from motion, facilitating scene rendering from novel perspectives. Moreover, we have developed a disentangled two-stage approach to 3D motion forecasting. This method initially focuses on camera ego-motion before addressing the residual motion of dynamic objects (e.g., cars, people), leading to more precise modeling of future scene motion.

Future Directions.

Longer-horizon future forecasting We aim to extend the predictive capabilities of our models to cover longer time spans, addressing the challenge of maintaining accuracy over extended future projections.

Forecasting occlusions for the future Our research will explore methods to accurately predict and manage occlusions in future frames, enhancing the model's ability to anticipate and navigate potential visibility issues.

0.3 Multimodal Future Forecasting

We aim to develop multimodal future forecasting models by integrating diverse cues, including vision and agent action intentions expressed through language. This integration significantly enhances the determinism of future forecasts beyond what visual input alone can achieve. Incorporating agent action intentions allows these models to provide more precise forecasts, which is particularly beneficial for the safety of autonomous driving systems. By considering multiple plausible agent and ego actions and foreseeing the future and obtaining feedback from different futures, before actual decision-making, can provide more rational planning, enhancing generalization and safety in end-to-end autonomous driving.

Contributions.

Language encoded agent intentions for future forecasting: Agent intentions often drive the actions within an environment, yet their significance has been underexplored in existing forecasting models. We propose to leverage diffusion models to create language-conditioned future forecasting models that incorporate language-based agent intentions alongside past video data. This approach introduces a novel mechanism of integrating explicit agent intentions into the forecasting process, offering a more nuanced and context-rich understanding of future states. hrough the development of sophisticated cross-modal interaction mechanisms and specialized loss functions, we ensure that the generated future videos accurately align with the specified agent intentions. This advancement marks a significant step towards achieving more comprehensive and reliable future forecasts, vital for navigating complex environments and enhancing the autonomy of systems.

Future Directions.

Multimodal forecasting with extended sensory Iiputs Enhancing our multimodal forecasting models will involve adding auditory cues and surface physical properties to our current vision and language inputs. Sound can offer insights into off-screen activities, while surface properties can inform on interaction dynamics. These additions aim to boost forecast precision and situational awareness, especially in complex environments like autonomous driving, by providing a more holistic understanding of the surroundings.

References

1. Wu, J., Yarram, S., Liang, H., Lan, T., Yuan, J., Eledath, J., Medioni, G.: Efficient video instance segmentation via tracklet query and proposal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 959–968 (2022)
2. Yarram, S., Wu, J., Ji, P., Xu, Y., Yuan, J.: Deformable vistr: Spatio temporal deformable attention for video instance segmentation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3303–3307. IEEE (2022)
3. Yarram, S., Yang, M., Yuan, J., Qiao, C.: Joint global-local alignment for domain adaptive semantic segmentation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3768–3772. IEEE (2022)
4. Yarram, S., Yuan, J., Yang, M.: Adversarial structured prediction for domain-adaptive semantic segmentation. Machine Vision and Applications **33**(5), 67 (2022)